Adversarial Combinatorial Bandits with General Non-linear Reward Functions

Yanjun Han (Stanford University)

Joint work with:

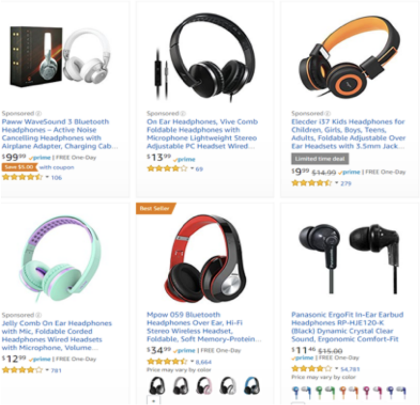
Xi Chen — New York University

Yining Wang — University of Florida

ICML 2021

Assortment optimization

- select a subset of substitutable items to maximize expected revenue
- recommendation in online retailing



Multinomial Logit model

mathematical model of assortment optimization:

- N available items in the pool
- each item has a revenue $r_i \in [0, 1]$, and a choice probability $v_i \in [0, 1]$
- seller offers an assortment $S \subseteq [N]$ of size K
- customer selects item i with probability

$$p_i(S, v) = \frac{v_i}{\underbrace{1}_{\text{“no-purchase”}} + \sum_{j \in S} v_j}$$

- seller's observation: the chosen item or “no-purchase”
- seller's expected revenue when offering assortment S:

$$R(S, v) = \sum_{i \in S} p_i(S, v) r_i = \frac{\sum_{j \in S} r_j v_j}{1 + \sum_{j \in S} v_j}$$

Static vs. dynamic assortment optimization

regret in repeated assortment optimization:

$$\mathbb{E}\left[\max_{S:|S|=K}\sum_{t=1}^{T}R(S,v_t)-\sum_{t=1}^{T}R(S_t,v_t)\right]$$

static model: $v_t \equiv v$ for all $t \in [T]$

- $\widetilde{O}(\sqrt{NT})$ regret achievable [Rusmevichientong et al. 2010, Agrawal et al. 2019, ...]

dynamic model: v_t may change across time

- open question: is $O(\sqrt{\mathrm{poly}(N,K)T})$ regret still achievable under the dynamic setting?

A more general combinatorial bandit

adversarial combinatorial bandit:

- time horizon T, number of arms N
- at each time $t \in [T]$:
  - a reward vector $v_t \in [0,1]^N$ is chosen
  - the learner chooses $S_t \subseteq [N]$ of size K, and observes bandit feedback

$$r_t \sim \text{Bernoulli}\left(R(S_t, v_t)\right), \quad \text{where } R(S_t, v_t) = g\left(\sum_{j \in S_t} v_{t,j}\right)$$

  - $g : \mathbb{R}_+ \to [0,1]$ is a known link function
- learner's regret:

$$\mathbb{E}\left[\max_{S:|S|=K} \sum_{t=1}^{T} R(S, v_t) - \sum_{t=1}^{T} R(S_t, v_t)\right]$$

assortment optimization with unit revenue: $g(x) = x/(1+x)$

Main result

Theorem

For general adversarial combinatorial bandits, the optimal regrets are:

- $\widetilde{\Theta}_{g,K}(\sqrt{TN^d})$ if g is a polynomial of degree $d \le K$;
- $\widetilde{\Theta}_{g,K}(\sqrt{TN^K})$ if g is not a polynomial of degree $\le K$.

implications:

- optimal regret crucially dictated by whether the link function is a low-degree polynomial or not
- since $g(x) = x/(1+x)$ is not a polynomial, $O(\sqrt{\text{poly}(N,K)T})$ regret is impossible in dynamic assortment selection

Proof idea

- consider assortment optimization with $K = 2$
- v_t drawn iid from the following distribution: choose $(i^\star, j^\star) \in \binom{[N]}{2}$ uniformly at random, and

$$v_k \equiv \frac{1}{2}, \quad k \notin \{i^\star, j^\star\}, \qquad (v_{i^\star}, v_{j^\star}) = \begin{cases} (1,1) & \text{w.p. } 1/4, \\ (0,1) & \text{w.p. } 3/8, \\ (1,0) & \text{w.p. } 3/8. \end{cases}$$

- key property: the multinomial distribution

$$\mathbb{E}\left(\frac{1}{1+v_i+v_j}, \frac{v_i}{1+v_i+v_j}, \frac{v_j}{1+v_i+v_j} \right)$$

is always $(1/2, 1/4, 1/4)$ unless the precise pair $(i^\star, j^\star)$ is chosen

- this type of construction is possible whenever g is not a low-degree polynomial, but requires involved real & functional analysis